Soraya Khan Junior MLOps Engineer

San Francisco, CA · soraya.khan@gmail.com · +1 415-555-0192 · linkedin.com/in/sorayakhan

PROFILE SUMMARY

- **Junior MLOps Engineer** with **2 years** on the ML platform team at **Scale AI**, after 2 years as a consultant at Bain, running **training pipelines and model deployments** for data quality models.
- Works in **Python** and Bash, builds pipelines in **Kubeflow Pipelines** and Airflow, tracks runs in **MLflow**, deploys with **Docker** and KServe on **Amazon EKS**, and watches models in Grafana.
- Knows **pipeline containerization**, **model registry promotion** and **data validation checks**, and is learning **GPU autoscaling** and drift monitoring from the senior engineers on the team.
- Works with **applied ML engineers** and the data team on the models that score labeling quality, and joins the **weekly deploy review** where new model versions get approved.
- Wrote the team's **pipeline onboarding guide**, answers platform questions in the support channel, and holds the **AWS Certified Machine Learning Engineer, Associate** certification.

TECHNICAL SKILLS

Languages & Scripting:	Python, Bash, SQL
ML Platforms & Tracking:	MLflow, Kubeflow, Amazon SageMaker (basics)
Pipelines & Orchestration:	Kubeflow Pipelines, Apache Airflow
Model Serving:	KServe, FastAPI, ONNX Runtime (learning)
Feature & Data Tools:	pandas, Great Expectations, Amazon S3, PostgreSQL
Cloud, Containers & CI/CD:	AWS, Amazon EKS, Docker, Helm, GitHub Actions, Terraform (basics)
Observability & Drift:	Prometheus, Grafana, Evidently (learning)

EDUCATION

Amazon Web Services — **AWS Certified Machine Learning Engineer, Associate** — Online · Mar 2025

University of California, Davis — **B.A. in Economics** — Davis, CA · Sep 2018 to Jun 2022

WORK EXPERIENCE

Scale AI — Junior MLOps Engineer — San Francisco, CA · Aug 2024 to Present

- Moved **9** labeling quality models from scripts on a shared VM into **Kubeflow Pipelines** with one container per step, so a retrain that took a full day of hand work now runs in **3 hours**.
- Added a **Great Expectations** check on every training set that stops the run when label counts or null rates look wrong, catching **6** broken data exports before training in my first year.
- Packaged models as **KServe** services on **Amazon EKS** GPU node pools with Helm charts and a GitHub Actions build, cutting the steps to deploy a new model version from 14 to **3**.
- Logged every training run and its dataset version in **MLflow** and set up registry stages with a reviewer sign-off, so **all 9** production models can be traced back to their data.
- Built **Grafana** dashboards for request latency, error rate and GPU memory per model, and wrote alert rules that cut the time to notice a failing endpoint from about a day to **15 minutes**.

Bain & Company — Associate Consultant — San Francisco, CA · Jul 2022 to Jul 2024

- Rebuilt a retail client's weekly demand forecast refresh in **Python** and Airflow, replacing a 40-tab spreadsheet and cutting the refresh from 2 days to **4 hours**.
- Wrote validation checks on **120** store data feeds that caught missing and duplicated sales files before the forecast ran, and logged each issue for the client's data team.
- Moved a pricing model from a laptop notebook into a scheduled **Docker** job on AWS with versioned inputs, so the client team could rerun any past week's result in **minutes**.